A Probability Method for Comparing Varieties against Checks

T. A. Jones

ABSTRACT

Combining results of tests conducted over a period of years is desirable when the tests are conducted at a single location or across locations known not to exhibit large nonrandom, predictable genotype × location interactions. But when cumulative reports across tests are attempted, complications arise because of unbalanced structure. The objective of this work was to derive a method to compare varieties against checks using accumulated data in the usual situation, where varieties are changing over the years the accumulated tests are conducted. The resultant probability method is easy to use, easy to present in an extension publication, and easy to interpret by the extension audience. It tests the null hypothesis that a variety equals a check in performance and permits calculation of the probability that the variety equals a check in performance when the null hypothesis is rejected (type 1 error). Calculations are simplified if data are coded and standardized for each test year. The method was used on an unbalanced data set yield of 23 varieties and 35 test years of alfalfa (*Medicago sativa* L.) grown near Ames, IA. Probability of one-tailed Type I error ranged from 0.47 to near 0.00 when data were coded on a test-year mean basis, and from 0.86 to near 0.00 when data were coded on the basis of the mean of 'Saranac' and 'Vernal'. The probability method determined that no more than seven test years were necessary to evaluate these varieties.

DUNNETT'S PROCEDURE (5) is often recommended for comparing all means with a control in a single experiment (13). Dunnett's critical value is tabulated for an experiment-wise error rate. However, because comparison with a check variety is often an objective in variety testing, a comparison-wise error rate is more appropriate. Combining results across tests conducted at a single location or across locations known not to exhibit significant amounts of nonrandom, predictable genotype × location (G × L) interaction is desirable. But this makes the data set unbalanced because typically varieties are included in many tests, but not all varieties are included in every test. Varieties included in tests change from year to year as new varieties become available and old varieties become obsolete (7).

Agricultural experiment stations in Illinois, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, Oklahoma, South Dakota, and Wisconsin participate in the Central Alfalfa Improvement Conference (CAIC). Of these 12 states (1–4, 6, 8–11, 14, 15, K.D. Kephart, 1987, personal communication), Iowa, Michigan, Nebraska, and North Dakota report results accumulated at individual locations since 1960, 1964, 1963, and 1959, respectively. Minnesota and Wisconsin report results accumulated over all locations since 1959 and 1982, respectively. Illinois, Kansas, Missouri, Ohio, Oklahoma, and South Dakota do not report cumulative results.

Current procedures used in reporting alfalfa yield test results emphasize comparison with a check (7). Of 12 states surveyed, seven express yield as a percentage of Vernal. Nebraska uses the mean of two checks, Missouri uses the mean of three checks, Illinois uses the mean of four checks, and Kansas and Oklahoma use the test mean. The CAIC has endorsed the use of Vernal, 'Baker', 'Riley', and 'Saranac AR' as checks, but only Illinois, Iowa, and North Dakota have followed this recommendation in recent years (J.B. Ogg, 1986, personal communication). Though expressing yield of a variety as a percentage of a single check is the most common procedure for combining data across tests, Hill and Rosenberger (7) judged it to exhibit the greatest prediction error of seven methods they compared. Their other methods included use of the predicted difference between varieties and checks, least-squares means (LSM) calculated in a two-way (varieties and tests) analysis of variance (ANOVA), LSMs calculated in a two-way ANOVA with each test weighted for its variance, and three variations of best linear unbiased prediction (BLUP). The BLUP methods exhibited the smallest prediction errors.

Hill and Rosenberger's (7) objective was to determine which of their methods for estimating mean yields from unbalanced, combined yield tests had the smallest prediction error. An alternate approach is to consider the null hypothesis that a variety equals the designated check or mean of checks in performance, based on the combined data. The probability that any given variety, in fact, equals the check when the null hypothesis is rejected is the Type I error. This approach retains the intuitive appeal of providing direct comparison with the check. The objective here was to develop a method to calculate this probability. It was then applied to 35 test years of unbalanced data collected from alfalfa yield tests near Ames, IA, conducted by the Iowa Agricultural and Home Economics Experiment Station.

MATERIALS AND METHODS

The tests were rotated among seven tile-drained fields in the Clarion-Nicollet-Webster soil association 13 km west of Ames, IA (10). These soils were classified in the order of mollisols and the suborder of aquolls. A new test was planted every spring and harvested the establishment year if growth and weed control permitted. Tests were subjected to a four cut-per-year management for the following 3 yr before being terminated. Thus, four tests were in rotation in any one year, one in the establishment year and three in production years. All tests included four replications of a lattice design. Tests were fertilized with P and K in accordance with soil test recommendations.

Total annual yield of production years was used in this study. Establishment year data were not considered. Twenty-three varieties were present in at least 12 of the 35 test years since 1972 and were included in this study. Two of the 23, Saranac and Vernal, were present in all 35 test years. Fifteen of the 23 were proprietary varieties and were assigned the letters A to O.

USDA-ARS, Forage and Range Res. Lab., Utah State Univ., Logan, UT 84322-6300. Joint contribution of USDA-ARS Forage and Range Res. Lab. and the Utah Agric. Exp. Stn., Journal Paper no. 3469. Received 21 Dec. 1987. *Corresponding author.

Published in Crop Sci. 28:907–912 (1988).

Before analysis, spreadsheet software coded and standardized data in each test year. To code data for test-year effect we used the test-year grand mean (Method 1) or mean of Saranac and Vernal, the two varieties included in every test year (Method 2). To standardize, coded data were divided by the standard error of a variety mean (SEM). The SEM was calculated from the least significant difference (LSD) of each test according to the formula

$$\text{SEM} = \text{LSD}(0.05)/\,(\text{t}(0.05) \times 1.414)\,.$$

Degrees of freedom (df) for error of the 35 test years ranged from 63 to 189.

Using appendix Eq. [5, 8, and 10], estimates of z for the null hypothesis that a variety was equal to Vernal were calculated as

$$z = \frac{\bar{y}_j - \bar{y}_{j'}}{\left\{\frac{1}{n_j^2}[n_j + \hat{R}_j n_j(n_j - 1)] + \frac{1}{n_{j'}^2}[n_{j'} + \hat{R}_{j'} n_{j'}(n_{j'} - 1)]\right\}^{1/2}},$$

where $\bar{y}_j$ is the mean of a variety over all n_j test years it is entered, $\bar{y}_{j'}$ is the mean of the check over all $n_{j'}$ test years, and $\hat{R}_j$ and $\hat{R}_{j'}$ are means of correlation coefficients between all test years including variety j and check j', respectively, weighted for degrees of freedom (appendix Eq. [8]).

Probabilities of Type I error corresponding to these z values were calculated from a table of normal curve areas as 0.5 − (area between 0 and z) for positive z and 0.5 + (area between 0 and z) for negative z.

RESULTS AND DISCUSSION

When calculated according to Method 1, on a test year mean basis, the probability of Type I error ranged from $P = 0.47$ for Saranac to nearly $P = 0.00$ for proprietary variety F (Table 1). When calculated according to method 2, on a Saranac/Vernal mean basis, the probability of Type I error ranged from $P = 0.86$ for 'Dawson' to nearly $P = 0.00$ for proprietary variety F (Table 2). Ranks of variety performance as determined by the two methods were correlated at $r = 0.92$ ($P < 0.01$). Most of the discrepancy between ranks was among the higher-yielding varieties. The P-values, as calculated in either method, have no known distribution themselves. They are intended for comparison of a variety against a check, not against every other variety.

The advantage of Method 2 relative to Method 1 is that the same two checks are used to code data for test-year effect in all test years. Thus, variety effects are not confounded with test-year effects. Using LSMs to code data for test-year effects also would successfully eliminate this problem. Method 1 suffered from the fact that the varieties tested over the years were improving in performance, increasing the test-year effect over time. Least-squares analysis indicated that yields increased 0.31 Mg ha^{-1} yr^{-1}. Thus, the test-year effect was confounded with genetic effects. Because varieties were not all tested in the same test years, in Method 1 varieties tested in earlier test years were at an advantage relative to varieties tested in later test years. This made Method 1 a conservative one for declaring newer varieties improved over older ones. As an example, because 'Kanza' and Dawson were entered only in earlier test years and Vernal was entered in all test years, Method 1 ranked Kanza and Dawson above Vernal while Method 2 ranked them below Vernal. The only advantage of Method 1 over Method 2 is that the influence of genotype × environment (G × E) interaction on the test-year effect, used for coding, is smaller in method 1 because of the larger number of varieties used to calculate it.

Rank of variety performance calculated by the standard method, as percentage of Vernal, was correlated with rank using Method 1 at $r = 0.88$ ($P < 0.01$) and with Method 2 at $r = 0.98$ ($P < 0.01$). The very high correlation of the standard method and Method 2 is partially due to their similarity, one correcting for performance of Vernal and the other correcting for performance of Vernal and Saranac, a variety which per-

Table 1. Ranking of 22 varieties in order of probability of failure to exceed Vernal (Type I error) on a test-year mean basis (method 1).

Variety	% of Vernal	Test years	Deviation from Vernal	Z	Probability Z ≤ 0
		no.	Mg ha^{-1}		
F	109	20	6.18	3.13	0.00
B	105	12	6.63	3.03	0.00
K	109	17	5.81	2.99	0.00
D	109	17	5.78	2.98	0.00
C	107	21	5.81	2.92	0.00
A	110	14	5.64	2.93	0.00
J	105	20	5.48	2.91	0.00
H	108	14	6.21	2.87	0.00
Riley	107	22	4.90	2.47	0.01
I	105	12	4.76	2.26	0.01
N	107	23	4.44	2.23	0.01
M	107	14	4.12	2.14	0.02
G	102	17	3.79	1.84	0.03
O	103	19	3.26	1.69	0.05
Saranac AR	105	20	2.83	1.49	0.07
Baker	104	25	2.03	1.01	0.16
E	102	16	1.57	0.80	0.21
L	104	14	1.06	0.55	0.29
Agate	100	26	1.03	0.51	0.31
Kanza	95	15	0.54	0.29	0.39
Dawson	96	15	0.42	0.23	0.41
Saranac	100	35	0.16	0.08	0.47
Vernal		35			

Table 2. Ranking of 22 varieties in order of probability of failure to exceed Vernal (Type I error) on a Saranac/Vernal mean basis (method 2).

Variety	% of Vernal	Test years	Deviation from Vernal	Z	Probability Z ≤ 0
		no.	Mg ha^{-1}		
F	109	20	8.02	4.06	0.00
A	110	14	7.73	4.01	0.00
K	109	17	7.65	3.94	0.00
D	109	17	7.62	3.92	0.00
H	108	14	7.86	3.68	0.00
C	107	21	6.92	3.48	0.00
Riley	107	22	6.45	3.25	0.00
N	107	23	6.03	3.02	0.00
M	107	14	5.84	3.02	0.00
B	105	12	6.31	2.88	0.00
I	105	12	5.52	2.62	0.00
J	105	20	4.11	2.18	0.01
Saranac AR	105	20	3.70	1.94	0.03
O	103	19	3.45	1.79	0.04
Baker	104	25	3.38	1.68	0.05
G	102	17	3.32	1.61	0.05
L	104	14	2.47	1.27	0.10
E	102	16	2.44	1.24	0.11
Agate	100	26	0.55	0.28	0.39
Saranac	100	35	0.16	0.08	0.47
Vernal		35			
Kanza	95	15	−1.92	−1.04	0.85
Dawson	96	15	−0.24	−1.10	0.86

formed similarly to Vernal. The difference in the correlation between the standard method and Method 1 (0.88) and the correlation between the standard method and Method 2 (0.98) is probably related to the greater confounding of genetic and test-year effects in Method 1 relative to Method 2. Here, with numbers of tests years 12 and greater, rank of varieties using the probability method was similar to rank using the standard method (Tables 1, 2). But unlike the standard method, the probability method considers the number of test years directly in the calculation of its parameters, which may be critical when conclusions are being drawn from only a few test years of data.

Method 2 is superior to Method 1 when testing of older varieties is being terminated as testing of newer varieties is beginning, because in Method 2 confounding of genetic effects with test-year effects is minimized. The objection to Method 2, the confounding of G × E interaction with test-year effects, would be minimized if more checks were included in every test. This would improve estimation of the coding terms, which are less precise in Method 2 because they are calculated with fewer observations than in Method 1. The CAIC recommends a standard set of four check varieties for inclusion in midwestern variety trials, which addresses this concern. These checks, however, are consistently used by only three of the 12 CAIC agricultural experiment stations. The checks are also updated every several years, so they do not remain the same over a long period of time as suggested here.

The probability method has three important practical attributes. First, the probability method is simple for calculation and presentation. This is especially important for a method to be used for extension purposes. Data are easily coded and standardized with values currently present in variety test reports. The number of test years of any variety j under test (n_j) and check ($n_{j'}$) and means of any variety j under test ($\bar{y}_j$) and check ($\bar{y}_{j'}$) are easily computed. Weighted means of correlation coefficients ($\hat{R}_j$ and $\hat{R}_{j'}$) from appendix Eq. [8] can be calculated by hand or with a simple computer program. Results can be presented in a single column of probabilities rather than presentation of separate results for each test which may not agree, confusing the reader.

Second, a straightforward policy, established in advance of testing, can be used to interpret test results. For example, varieties would be identified as superior to the check only when enough tests were conducted and paid for, in the case of proprietary varieties, to show that the probability of Type I error had fallen below a certain threshold, e.g., 0.05 or 0.01. Only then would the variety be recommended over the check. Extension publications would state that use of varieties not tested sufficiently for the error to fall below the threshold could not be considered better than the check. An advantage of the probability method is that it provides a well defined cutoff point for making this decision, rather than suggesting care by the reader before choice of any variety with data from only a few test years.

Producers are most familiar with varietal ranking based on percentage of a check. Though aware that such rankings are influenced by variable numbers of test years among varieties and variable precision of tests, producers commonly use rankings verbatim. An extension effort will be needed to educate producers about the ability of the probability method to deal with these problems. Though the Type I error concept will be new to most in the extension audience, it is easily understood when accompanied with an example. Obviously superior varieties will soon pass the probability threshold for recommendation, while new varieties marginally better than the check will require additional testing before recommendation. Others, no better than the check, will fail to decrease in probability of Type I error and will not be recommended over the check.

Third, the probability method is able to determine appropriate sample sizes. Tests should be continued until the response to the null hypothesis, that a variety equals the check in performance, is clear. This assures that enough testing has been done to settle the question, at least at the predetermined threshold level of probability. The method also allows waste of money and resources, because of continued testing after the threshold probability has been reached, to be eliminated. This is a boon to researchers burdened with needlessly large numbers of entries and to companies charged for unnecessary testing. In this experiment weighted means of correlation coefficients averaged 0.37. Using appendix Eq. [10], it was clear that testing beyond seven test years was redundant (Fig. 1). Thus, it is reasonable that when number of test years is well above this number, as in this experiment, the standard method gives similar results to method 2 ($r = 0.98$). A good estimate of the average correlation coefficient and the range that can be expected under normal conditions should be acquired for any testing program using the probability method.

The probability method is appropriate for data combined with the intention of making recommendations for a particular geographical area. The model used in the probability method does not include a G × E interaction term and makes no assumptions regarding G × E interaction effects. For this reason, test years exhibiting nonrandom, predictable G × L interactions should not be combined. Combining data

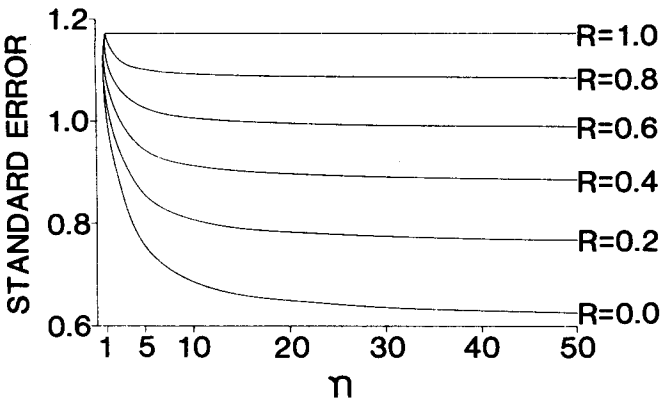

Fig. 1. Standard error of the difference between yield performance of Vernal and a variety under test calculated from appendix Eq. [10]. Numbers of test years (n) and various weighted means of correlation coefficients among test years (R) are indicated. Variance (Vernal) = 0.372.

from test years exhibiting genotype × year interaction, however, is within the limits of the probability method, because the environment of any year is random, or at least unpredictable. To avoid bias in interpretation, the test years of the data set must be a sample of the population of test years to which inference will be made. Scientists conducting variety tests have long avoided combining data from locations with distinct climates or soils because confounding large amounts of G × E interaction is likely. This need is not lessened by this or any other statistical procedure. Cluster analysis can be used to group together locations with a predetermined maximal degree of heterogeneity.

Any two test years have a correlation coefficient less than 1.0 because of experimental error and G × E interaction. The probability method acknowledges this effect of the correlations among test years on the denominator of the z statistic. In contrast, a conventional LSD, not designed for combined data, assumes independence of all observations and thus does not consider correlations. Test years may be correlated to varying degrees because of similarities in climate, weather, or soils. Inclusion of the correlation coefficients increases the denominator of the z statistic by increasing the standard error (appendix Eq. [5]). The lower the correlation coefficients among test years, the more useful additional test years are in lowering the standard error (Fig. 1). If the correlation coefficients all equal 1.0, no reduction in standard error results from additional testing because no additional information is generated. Increases in the denominator because of increased n are exactly balanced by increases in the numerator because of increased sums of correlation coefficients. For the 23 varieties used here, the weighted mean of correlation coefficients averaged 0.37, with a range extending from 0.26 for Dawson and Kanza to 0.55 for variety B. At $r = 0.37$ little reduction in standard error was apparent after the seventh test year of data (Fig. 1). Even at $r = 0$ little reduction in standard error was apparent after the tenth test year of data.

Whereas the intention of the methods described by Hill and Rosenberger (7) was to improve estimates of variety means, the intention of the probability method is to estimate the probability of the Type I error associated with the rejection of the null hypothesis. Thus, these two methods are not mutually exclusive and combining them into one procedure may be desirable in a research context, where need for simplicity is of secondary importance.

Appendix

Model

Let

$$Y_{ij} = u + t_i + v_j + e_{ij},$$

where

Y_{ij} = mean of all replicates of annual yield for variety j in test year i,
u = fixed overall mean,
t_i = fixed effect of test year i, as described in Materials and Methods,
v_j = random effect of variety j, and
e_{ij} = random measurement error of variety j in test year i, including any interaction between t_i and v_j.

If all Y_{ij}'s are coded by subtracting $(\hat{u} + \hat{t}_i)$ and standardized by dividing by $\hat{s}_i$, the estimated standard error of Y_{ij}, then

$$\hat{y}_{ij} = \frac{Y_{ij} - \hat{u} - \hat{t}_i}{\hat{s}_i} = \frac{\hat{v}_j + \hat{e}_{ij}}{\hat{s}_i}. \qquad [1]$$

Assumptions

The theory presented below assumes that the e_{ij}'s in a test year are independently and identically normally distributed with mean = 0 and variance = σ_e^2. Thus,

$$\text{cov}\,(e_{ij}, e_{ij'}) = 0. \qquad [2]$$

Also, the variety effects must be independently and identically normally distributed with mean = 0 and variance = σ_v^2.

In addition, varieties j under test must be independent of check j', which is fulfilled when plots are properly randomized within and between test years. Thus,

$$\text{cov}\,(v_j, v_{j'}) = 0. \qquad [3]$$

Also, variety effects must be independent of the e_{ij}'s. Thus,

$$\text{cov}\,(v_j, e_{ij}) = \text{cov}\,(v_j, e_{ij'}) = 0. \qquad [4]$$

Test Statistic

The difference between two standardized treatment means, $\bar{y}_{\cdot j}$ and $\bar{y}_{\cdot j'}$, divided by its estimated standard error, approximately conforms to a z distribution. The hypothesis that the true standardized treatment means are equal may be tested with a one-tailed z test:

$$\frac{\hat{\bar{y}}_{\cdot j} - \hat{\bar{y}}_{\cdot j'}}{\hat{\sigma}(\bar{y}_{\cdot j} - \bar{y}_{\cdot j'})} \approx z. \qquad [5]$$

The degrees of freedom of this quantity will always be large in practice because they are derived from the degrees of freedom of the s_i's used to standardize the data. After even a single test year of a variety testing program, enough degrees of freedom are available to justify the use of a z distribution instead of a t distribution. The quantity in [5] is distributed only approximately as z, not only because of a finite number of degrees of freedom, but also because of errors in the estimation of values needed for its calculation. The coding and standardizing terms [1], used in the numerator of [5], and the correlation coefficients among test years, to be used for calculation of the denominator of [5], are estimated with error.

The standard error in the denominator of [5] is the square root of the variance, where

$$\text{var}\,(\bar{y}_{\cdot j} - \bar{y}_{\cdot j'}) \doteq \text{var}\,(\bar{y}_{\cdot j}) + \text{var}\,(\bar{y}_{\cdot j'}) - 2\text{cov}\,(\bar{y}_{\cdot j}, \bar{y}_{\cdot j'}) \qquad [6]$$

Variance of a Mean

The variance of a mean can be expressed as

$$\text{var}\,(\bar{y}_{\cdot j}) \doteq \text{var}\left[\frac{(y_{1j} + y_{2j} + \ldots + y_{n_j j})}{n_j}\right] =$$

$$\frac{1}{n_j^2}\left[\sum_{i=1}^{n_j} \text{var}\,(y_{ij}) + \sum_{i=1}^{n_j}\sum_{\substack{i'=1 \\ i \neq i'}}^{n_j} \text{cov}\,(y_{ij}, y_{i'j})\right],$$

and likewise for var $(\bar{y}_{\cdot j'})$. But because data are standardized,

$$\text{var}\,(y_{1j}) \doteq \text{var}\,(y_{2j}) \doteq \text{var}\,(y_{ij}) \doteq 1\,,$$

where these are the diagonal elements of the correlation matrix of the n_j test years for variety j. Now, it is well known that a correlation coefficient is the covariance of standardized variables (12). Thus,

$$\text{cov}\,(y_{ij}, y_{i'j}) \doteq r_{ii'}$$

the correlation coefficient between test years i and i', where each $r_{ii'}$ $(i \neq i')$ is an off-diagonal element of the correlation matrix. So if data are standardized,

$$\text{var}\,(\bar{y}_{\cdot j}) \doteq \frac{1}{n_j^2}\,(n_j + \sum_{\substack{i=1 \\ i \neq i' \\ i,\, i' \text{ including } j}}^{n_j} \sum_{i'=1}^{n_j} r_{ii'})\,, \qquad [7]$$

where the summations are over pairs of test years i and i', both including variety j, and likewise for var $(\bar{y}_{\cdot j'})$.

Because

$$r_{ii'} \doteq \text{cov}\,(y_{ij}, y_{i'j}) = \text{cov}\left[\frac{v_j + e_{ij}}{s_i}, \frac{v_j + e_{i'j}}{s_{i'}}\right]$$

$$= \frac{1}{s_i s_{i'}}\,[\text{var}\,(v_j) + \text{cov}\,(v_j, e_{i'j}) + \text{cov}\,(e_{ij}, v_j) + \text{cov}\,(e_{ij}, e_{i'j})]$$

$$= \frac{\sigma_v^2 + \text{cov}\,(e_{ij}, e_{i'j})}{s_i s_{i'}} \text{ by [1] and [4])},$$

the correlation between test years i and i' consists of one term corresponding to the variance of the variety effect and another term corresponding to the covariance between error effects in test years i and i'. The cov $(e_{ij}, e_{i'j}) <$ var $(e_{ij}) = \sigma_e^2$ when G × E interaction is present because the interaction is included in the e_{ij} term in the model. Thus, the correlation coefficient between test years exhibiting no G × E interaction will be higher than that between test years i and i' exhibiting G × E interaction.

If the correlation coefficients are based on different degrees of freedom, it is appropriate to weight them by their individual degrees of freedom, similar to weighted variances (12). In this case, in Eq. [7]

$$\sum_{\substack{i=1 \\ i \neq i' \\ i,i' \text{ including } j}}^{n_j} \sum_{i'=1}^{n_j} r_{ii'} \text{ is replaced by}$$

$$\left(\frac{\sum_{\substack{i=1 \\ i \neq i' \\ i,i' \text{ including } j}}^{n_j} \sum_{i'=1}^{n_j} r_{ii'}\,(\text{df}_{r_{ii'}})}{\sum_{\substack{i=1 \\ i \neq i' \\ i,i' \text{ including } j}}^{n_j} \sum_{i'=1}^{n_j} \text{df}_{r_{ii'}}}\right) n_j(n_j - 1) = \text{R}_j\text{n}_j(\text{n}_j - 1)\,, \qquad [8]$$

where degrees of freedom (df) of $r_{ii'}$ equal the number of varieties common to test years i and $i' - 2$, and R_j equals the mean of correlation coefficients between all test years including variety j and check j', weighted for df.

Covariance Between Two Means

The covariance between the standardized treatment means of variety j under test and check j' can be expressed as cov $(\bar{y}_{\cdot j}, \bar{y}_{\cdot j'}) =$

$$\text{cov}\left[\frac{(y_{1j} + y_{2j} + \ldots + y_{n_j j})}{n_j}, \frac{(y_{1j'} + y_{2j'} + \ldots + y_{n_{j'} j'})}{n_{j'}}\right] =$$

$$\frac{1}{n_j n_{j'}}\left[\sum_{\substack{i=1 \\ i=i'}}^{n_j} \sum_{i'=1}^{n_{j'}} \text{cov}\,(y_{ij}, y_{i'j'}) + \sum_{\substack{i=1 \\ i \neq i'}}^{n_j} \sum_{i'=1}^{n_{j'}} \text{cov}\,(y_{ij}, y_{i'j'})\right].$$

But because data are coded, yes $i = i'$,

$$\text{cov}\,(y_{ij}, y_{i'j'}) \doteq \frac{1}{s_i}\,[\text{cov}\,(v_j + e_{ij}, v_{j'} + e_{ij'})]$$

$$= \frac{1}{s_i}\,[\text{cov}\,(v_j, v_{j'}) + \text{cov}\,(v_j, e_{ij'}) + \text{cov}\,(e_{ij}, v_{j'}) + \text{cov}\,(e_{ij}, e_{ij'})]$$

$$= 0$$

by Eq. [2, 3, 4].

So,

$$\sum_{\substack{i=1 \\ i=i'}}^{n_j} \sum_{i'=1}^{n_{j'}} \text{cov}\,(y_{ij}, y_{i'j'}) \doteq 0, \text{ and}$$

$$\sum_{\substack{i=1 \\ i \neq i'}}^{n_j} \sum_{i'=1}^{n_{j'}} \text{cov}\,(y_{ij}, y_{i'j'}) = 0, \text{ likewise.}$$

For coded data, then,

$$\text{cov}\,(\bar{y}_{\cdot j}, \bar{y}_{\cdot j'}) \doteq 0\,. \qquad [9]$$

Variance of the Difference Between Two Means

For coded and standardized data, then, by Eq. [6, 7, 9], var $(\bar{y}_{\cdot j} - \bar{y}_{\cdot j'}) =$

$$\frac{1}{n_j^2}(n_j + \sum_{\substack{i=1 \\ i \neq i' \\ i,i' \text{ including } j}}^{n_j} \sum_{i'=1}^{n_j} r_{ii'}) + \frac{1}{n_{j'}^2}(n_{j'} + \sum_{\substack{i=1 \\ i \neq i' \\ i,i' \text{ including } j'}}^{n_{j'}} \sum_{i'=1}^{n_{j'}} r_{ii'})\,. \qquad [10]$$

Eq. [10] is used to calculate the denominator of z in Eq. [5]. For

$$H_o\colon \bar{y}_{\cdot j} = \bar{y}_{\cdot j'}, \text{ and}$$
$$H_a\colon \bar{y}_{\cdot j} > \bar{y}_{\cdot j'}\,,$$

the value of z is associated with a probability that $\bar{y}_{\cdot j}$ fails to exceed $\bar{y}_{\cdot j'}$ when the null hypothesis is rejected (Type I error) for a one-tailed test. This is the probability that is used to compare varieties against checks in the probability method (Tables 1, 2).

Reduction to a Familiar Expression

It is easily seen that Eq. [10] simplifies to a familiar expression for the variance of the difference between two means in the balanced case when all correlation coefficients = 0. For example, if variety means among test years are independent, an assumption made by the LSD but not the probability method, Eq. [10] reduces to

$$\text{var}\ (\bar{y}_{\cdot j} - \bar{y}_{\cdot j'}) \doteq \frac{1}{n_j} + \frac{1}{n_{j'}}.$$

Then, if the data are not standardized and variances of test years are assumed equal,

$$\text{var}\ (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot j'}) = \frac{\sigma^2}{n_j} + \frac{\sigma^2}{n_{j'}}.$$

Then, if $n_j = n_{j'} = n$,

$$\text{var}\ (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot j'}) = \frac{2\sigma^2}{n},$$

the quantity used to represent variance of the difference in two means when calculating the LSD.

ACKNOWLEDGMENTS

Appreciation is extended to K. Roozeboom and I.T. Carlson for providing copies of Iowa alfalfa yield test reports. Suggestions made by D.V. Sisson and R.L. Canfield, professors, Dep. of Mathematics and Statistics, Utah State Univ., regarding the appendix of the manuscript, are also greatly appreciated.

REFERENCES

1. Anderson, B., R.S. Moomaw, R.W. Elmore, and P.E. Reece. 1985. Alfalfa variety selection. NebGuide G77–357. Cooperative Ext. Serv., Institute of Agriculture and Natural Resources, Univ. Nebraska, Lincoln, NE.
2. Barnes, D.K., and N.P. Martin. 1987. Forage crops. p. 3–9. *In* A.G. Lewis (ed.) Varietal trials of farm crops. Minnesota Report 24–1987. Agric. Exp. Station, Univ. Minnesota, St. Paul, MN.
3. Caddel, J.L., J.B. Ogg, K.T. Shelton, and L.M. Rommann. 1987. Alfalfa variety and strain evaluation in Oklahoma: 1973–1986. Oklahoma Ag. Exp. Stn. Res. Rep. P-895.
4. Dodds, D.L., and D.W. Meyer. 1987. Alfalfa variety selection. Cooperative Ext. Serv., North Dakota State Univ., Fargo, ND.
5. Dunnett, C.W. 1955. A multiple comparisons procedure for comparing several treatments with a control. J. Amer. Stat. Assoc. 50:1096–1121.
6. Graffis, D.W., D.A. Miller, D.M. Griffith, D.C. Faulkner, S.G. Carmer, and R.D. Seif. 1985. Forage crops variety trials in Illinois—1985. Dep. of Agronomy, Agric. Exp. Stn., Univ. Illinois, Urbana-Champaign, IL.
7. Hill, R.R., Jr., and J.L. Rosenberger. 1985. Methods for combining data from germplasm evaluation trials. Crop Sci. 25:467–470.
8. Minor, H.C., C.G. Morris, H.L. Mason, R.E. Mattis, D.R. Knerr, and E.E. Lawman. 1986. Forages, 1986 Missouri crop performance. Spec. Rep. 351. Agric. Exp. Stn., Univ. Missouri, Columbia, MO.
9. Rohweder, D.A., C.R. Grau, G.R. Blomberg, M. Culik, T. Drendel, J. Fanta, R. Ipson, J.H. Key, D. Nehring, W.H. Paulson, R.E. Rand, T.D. Syverud, G.C. Weis, and D. Wiersma. 1987. Perennial legume forage crop variety update for Wisconsin. Cooperative Ext. Serv., Univ. Wisconsin, Madison, WI.
10. Roozeboom, K., and I.T. Carlson. 1987. 1986 Iowa alfalfa yield test report. AG 84–6. Cooperative Ext. Serv., Iowa State Univ., Ames, IA.
11. Shroyer, J.P., E.L. Sorensen, T.L. Walter, R.J. Raney, M. Witt, R. Greenland, G.R. TenEyck, J.L. Moyer, and K. Janssen. 1987. Kansas alfalfa variety tests. 1986. Cooperative Ext. Serv., Kansas State Univ., Manhattan, KS.
12. Sokal, R.F., and F.J. Rohlf. 1969. Biometry—The principles and practice of statistics in biological research. W.H. Freeman and Co., San Francisco, CA.
13. Steel, R.G.D., and J.H. Torrie. 1980. Principles and procedures of statistics—A biometrical approach. 2nd ed. McGraw-Hill, New York.
14. Tesar, M.B., and R.H. Leep. 1986. Alfalfa variety recommendations for Michigan. Ext. Bull. E-1098. Cooperative Ext. Serv., Michigan State Univ., East Lansing, MI.
15. van Keuren, R.W., and D.K. Meyers. 1985. Ohio forage report. Agronomy Dep. Series 195. Ohio Agric. Res. and Development Ctr., The Ohio State Univ., Wooster, OH.

LETTER TO THE EDITOR

Comment on a Probability Method for Comparing Test Varieties with a Check

Evaluating variety tests accumulated in consecutive years, where varieties may change over the years, is a frequent task. A common approach is to compare each variety with a check. One is often faced with the problem that the check is grown in more years than the test variety. For the comparison to be meaningful it is desirable to use all information available, which includes those test years in which the check was grown, but during which the test variety under consideration was not grown. Statistical difficulties in the analysis of such data arise because they are unbalanced. Jones (1988) proposed a method for unbalanced data recorded at a single location, which yields the probability that a given variety is not better than the check. The procedure is based on the Type I error of a test of the null hypothesis that the variety equals the check. A major drawback of this method is that it leads to biased probability estimates in the presence of variety × year interactions. In this letter, two simple alternatives, i.e., the paired *t*-test and weighted least squares analysis, are offered which are appropriate in case of uncorrelated errors and interactions. It will be argued that Jones' approach is not relevant in most practical circumstances because the number of years required is prohibitive, and because plant breeders usually test promising cultivars at more than just one location.

Problems

The following are statistical problems with Jones' approach.

1. Genotype × year interaction is not appropriately considered.
2. A genotypic variance component is used, where it should not be present.
3. Jones assumes genotypes are random and years fixed, which is quite unusual. Most researchers will regard years (environments) as random, while assuming fixed genotypes (Shukla, 1972).

Also, the procedure is rather complicated because it assumes that errors (including random interactions) of different years are correlated. For the moment, we will adhere to Jones' assumptions and discuss the errors listed in Problems 1 and 2. Afterwards, a modification of Jones' method is suggested which assumes fixed genotypes and random environments (this addresses Problem 3). We drop the assumption of correlated errors and discuss alternative procedures under this simplified model.

To outline Problems 1 and 2, we take a glance at the model used by Jones (1988) in which

$$X_{ij} = u + t_i + v_j + e_{ij}, \qquad [1]$$

where X_{ij} = mean of all replicates of annual yield for variety j in test year i,

u = fixed overall mean,
t_i = fixed effect of test year i,
v_j = random effect of variety j,

and e_{ij} = random measurement error of variety j in test year i, including any interaction between year i and variety j. All random effects are taken to be normally distributed. It is assumed that cov (e_{ij}, e_{if}) = cov (v_j, v_f) = cov (v_j, e_{ij}) = cov (v_j, e_{if}) = 0.

Jones (1988) points out that test years may be correlated to varying degrees because of similarities in climate, weather, or soil; this is certainly a point to be considered. Therefore, cov $(e_{ij}, e_{i'j})$ is allowed to differ from zero. Jones' method is rather complicated mainly because of this assumption. Matters simplify considerably, if we assume that cov $(e_{ij}, e_{i'j})$ = 0. This simple case will be discussed later.

For a comparison of two varieties, Jones (1988) suggests the statistic z given by

$$z = \frac{\bar{y}_{.j} - \bar{y}_{.c}}{[\mathrm{var}(\bar{y}_{.j} - \bar{y}_{.c})]^{1/2}}.$$

Here, y_{ij} and y_{ic} are standardized values of observations X_{ij} and X_{ic}, the yields of varieties j and c in year i (see below). Means $\bar{y}_{.j}$ and $\bar{y}_{.c}$ are computed across all years in which the corresponding variety was grown. Regarding z as a standard normal deviate, the probability of finding a z-value larger than the observed z when the jth cultivar is equal to the check c, is computed from a table of the normal distribution. If this probability is smaller than 5%, it is concluded that the jth cultivar is in fact better than the check. The procedure is equivalent to a test of the null hypothesis that the jth cultivar is not better than the check. X_{ij}s are standardized by subtracting the test year grand mean (Method 1), or the mean of two check varieties (Method 2), and dividing by the standard error of a variety mean (s_i = SEM). The SEM is calculated from the least significant difference (LSD) of each test year according to the formula SEM = LSD(0.05)/ [t(0.05) × 1.414].

The aim of the standardization is to make y_{ij} a random deviate with unit variance. This is not achieved if there are variety × year interactions. The reason is that s_i is estimated from the residual mean square of the corresponding test year. This mean square is an estimate of the error variance of that year. It does not include the (random) variety × year interaction, which is confounded with variety effects in the variety means of a test year. For a comparison of variety means over several years, we need to consider all random variability across years, which is composed of interaction and error effects.

To clarify the point, consider the model used by Jones (1988). The random deviation e_{ij} in Eq. [1] may be expressed as

$$e_{ij} = tv_{ij} + \varepsilon_{ij},$$

where

tv_{ij} = interaction of variety j with year i and
ε_{ij} = mean error of variety j in year i.

The complete model for observations X_{ij} becomes

$$X_{ij} = u + t_i + v_j + tv_{ij} + \varepsilon_{ij}. \qquad [2]$$

For standardization, Jones (1988) suggests substracting from X_{ij} an estimate of $u + t_i$ (obtained by either Method 1 or Method 2) and dividing by s_i, the standard error of a mean in test year i. Here, we will use population values instead of estimates and compute

$$y_{ij} = \frac{X_{ij} - u - t_i}{\sigma_i}.$$

We have

$$X_{ij} - (u + t_i) = v_j + tv_{ij} + \varepsilon_{ij} \equiv Z_{ij}.$$

For a given genotype j, Z_{ij} has variance

$$\sigma_i^2 = \text{var}(Z_{ij}) = \text{var}\,(tv_{ij}) + \text{var}\,(\varepsilon_{ij}) = \sigma_{tv}^2 + \sigma_{\varepsilon i}^2,$$

where $\sigma_{tv}^2 = \text{var}\,(tv_{ij})$ is the year × variety interaction variance and $\sigma_{\varepsilon i}^2 = \text{var}\,(\varepsilon_{ij})$ is the error of a variety mean in year i. The standardized values $y_{ij} = Z_{ij}/\sigma_i = (X_{ij} - u - t_i)/\sigma_i$ are normally distributed with unit variance. To put it clearly, this statement refers to the variability of y_{ij}s measured in different years. For practical purposes u, t_i, σ_i^2 could be replaced by appropriate estimates as suggested by Jones (1988). In the presence of interaction, however, $s_i^2 = (\text{SEM})^2$ is not an estimator of σ_i^2, since it only estimates $\sigma_{\varepsilon i}^2$. It underestimates the variability of Z_{ij} across years, because interaction is not taken account of. Therefore, the standardization proposed by Jones (1988) leads to y_{ij} values which have a variance larger than unity. As a result, the estimated probabilities of exceeding the observed z-values will be in error.

Another problem is that Jones (1988) assumes cov $(y_{ij}, y_{i'j}) = [\sigma_v^2 + \text{cov}\,(e_{ij}, e_{i'j})]/\sigma_i\sigma_{i'}$, where σ_v^2 is the variance of random variety effects v_j. This is not correct, since we consider the variance of the mean (taken across years i) of one single variety j, which is independent of the variance σ_i^2 of varieties included in the analysis. Clearly, for a given variety j, the effect v_j in the expression $y_{ij} = (v_j + e_{ij})/\sigma_i$ does not change with i, and hence σ_v^2 should not be present in the expression for cov $(y_{ij}, y_{i'j})$.

Alternatives

I was unable to modify Jones' procedure, keeping the assumption of correlated errors and regarding genotypes as fixed and environments as random. It appears that a modification would require additional constraints on the correlation structure, which is common in time series analyses. One may take the simplified view that similarities between years are reflected by similar year effects t_i and need not be modelled by a covariance between e_{ij}s. It would be useful to test the null hypothesis that cov $(e_{ij}, e_{i'j})$ equals zero for every i,i', but this author is not aware of any procedure for this problem.

If we assume cov $(e_{ij}, e_{i'j}) = 0$ for all i,i', and $\sigma_i^2 = \sigma^2$ for all i as is usually done, the variance of a difference becomes

$$\text{var}(\bar{X}_{.j} - \bar{X}_{.c}) = \frac{\sigma_i^2 + \sigma^2}{n_j} + \frac{\sigma_i^2 + \sigma^2}{n_c} - \frac{2n_{jc}\sigma_i^2}{n_j n_c}, \quad [3]$$

where σ_i^2 is the variance of t_i, n_j (n_c) is the number of test years for variety j (c), and n_{jc} is the number of years in which variety j and c were tested together. (It is observed that for $n_j = n_c = n_{jc}$ this simplifies to the variance for the paired t-test.) With uncorrelated errors, σ_i^2 and σ^2 may be estimated by standard procedures (Searle et al., 1992). SAS PROC VARCOMP and PROC MIXED provide such estimates (SAS Institute, 1989 and 1992). These estimates are then used to estimate

$$z^* = (\bar{X}_{.j} - \bar{X}_{.c})/\sqrt{\text{var}(\bar{X}_{.j} - \bar{X}_{.c})}.$$

The estimate approximately follows a standard normal distribution. It may be used to test equality of variety j and check c.

SAS PROC MIXED may be used to obtain a weighted least squares analysis. One would analyse all available means X_{ij}. The following SAS code will do the appropriate computations:

```
PROC MIXED;
CLASS CULTIVAR YEAR;
MODEL YIELD = CULTIVAR;
RANDOM YEAR;
RUN;
```

By default, SAS computes REML (Restricted Maximum Likelihood) estimates of the variance components, but the METHOD option allows one to obtain ML (Maximum Likelihood) estimates and MIVQUEs (Minimum Variance Quadratic Unbiased Estimates) as well (SAS Institute, 1992). Differences among individual cultivars may be tested using the CONTRAST statement.

The weighted least squares approach is problematic, if only few degrees of freedom are available for estimating σ_i^2, i.e., if only few test years are available. It may be a better procedure to compute the paired t-test, with only those years during which the two genotypes to be compared were grown together (Bradley et al., 1988). The paired t-test discards the yield data of years in which the test variety was not grown, so not all information is exploited. In the example given by Jones (1988) this would mean discarding between 9 and 23 of 35 yr. The main advantage is that one does not need to estimate σ_i^2. Moreover we may relax the assumptions regarding the interaction plus error effects. So far we have assumed that $\sigma^2 = \text{var}\,(tv_{ij}) + \text{var}\,(\varepsilon_{ij}) = \sigma_{tv}^2 + \sigma_\varepsilon^2$ is equal for all varieties. It is often observed, however, that the interaction variance var (tv_{ij}) varies among varieties. This means that the variance $\sigma_j^2 = \text{var}(tv_{ij}) + \sigma_\varepsilon^2$ is not necessarily constant for all varieties. In fact, σ_j^2 is often considered a measure of phenotypic stability of variety j (Shukla 1972; Lin et al. 1986), and numerous investigations have revealed a heterogeneity in the σ_j^2 s (Kang and Miller 1984; Gravois et al. 1990). It is noted that heterogeneity of the variances σ_j^2 and σ_ε^2 does not invalidate the paired t-test.

All procedures described here are also applicable if a variety is to be compared to the mean of several checks. In the computations for the paired t-test one simply has to replace X_{ic} by the mean of checks in year i. Using PROC MIXED one has to define an appropriate contrast including several checks. For three checks, e.g., the coefficients would be 3 for the variety under consideration and -1 for each of the checks.

In summary, this author would prefer the paired t-test in case σ_j^2s are very heterogeneous and/or only a few test years are available. Otherwise, the weighted least squares procedure is probably more efficient.

The focus of Jones' approach is on identifying the best genotypes for a particular location and assessing variability across years. For this purpose, only yield data gathered at that particular site are of interest. A general problem, however, is that the selection of the best genotypes is typically made after 2 or 3 yr of testing, while Jones' approach requires many more years in order to obtain reasonably accurate results. For good estimates of σ_i^2 and σ^2, at least 10 (preferably more) years of testing are necessary. The same problem exists for the paired t-test and the weighted least squares approach. By the time enough data is gathered, the new cultivar will probably be obsolete.

Most plant breeders are not interested in particular environments but rather seek genotypes with broad adaptability. Promising genotypes are therefore usually tested in many locations, and recently there has been a tendency to reduce replications per location in favor of an increased number of locations (Bradley et al., 1988). In this case, plenty of data are available for head-to-head comparisons across environments by the paired t-test, and only 2 or 3 yr of testing are required.

Received 1 July 1993.

Faculty of Agricultural and Environmental Sciences
University of Kassel
FB 11 Steinstrasse 19
37213 Witzenhausen 1, Germany
(e-mail: piepho@wiz.uni-kassel.de)

H.-P. PIEPHO

References

Bradley, J.P., K.H. Knittle, and A.F. Troyer. 1988. Statistical methods in seed corn product selection. J. Prod. Agric. 1:34–38.

Gravois, K.A., K.A.K. Moldenhauer, P.C. Rohman. 1990. Genotype-by-environment interaction for rice yield and identification of stable, high yielding varieties. p. 181–188. *In* M. S. Kang (ed.) Genotype-by-environment interaction in plant breeding. Louisiana State University, Baton Rouge.

Jones, T.A. 1988. A probability method for comparing varieties against checks. Crop Sci. 28:907–912.

Kang, M.S., and J.D. Miller. 1984. Genotype × environment interaction for cane and sugar yield and their implications in sugarcane breeding. Crop Sci. 24:435–440.

Lin, C.S., M.R. Binns, and L.P. Levkovitch. 1986. Stability analysis: where do we stand? Crop Sci. 26:894–900.

SAS Institute. 1989. SAS/STAT users' guide, Version 6, 4th ed. Vol. 2. SAS Institute, Cary, NC.

SAS Institute. 1992. SAS/STAT software: changes and enhancements, Release 6.07. SAS Technical Report P-229. SAS Institute, Cary, NC.

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance components. Wiley, New York.

Shukla, G.K. 1972. Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29:237–245.

A Response to the Letter to the Editor from H.-P. Piepho

H.-P. Piepho has pointed out an error of importance in the Jones (1988) paper regarding standardization of data. This error is easily corrected by (i) standardizing data with the standard error calculated from the combined data set rather than the standard error of individual test years and (ii) assuming homogeneous error variances among test years. The procedure is referred to as the *revised probability method* for comparing varieties against checks. This response presents the revised probability method (model and assumptions). This is followed by comments regarding (i) a misunderstanding concerning the use of a covariance term in the probability method, (ii) the analogous nature of Piepho's Eq. [3] to [10] of Jones (1988), and (iii) the unintended use of the probability method to compare breeder's lines, a research activity, rather than the intended use to compare varieties against checks in variety testing as traditionally practiced in the USA, an extension activity. Finally, the procedure derived by Piepho in his letter, the paired *t* test advocated by Bradley et al. (1988), and the revised probability method, are discussed in the context of the extension activity of variety testing.

Revised Probability Method

Model

Let $Y_{ij} = \mu + t_i + v_j + e_{ij}$,

where

Y_{ij} = mean of all replicates of annual yield for variety j in test year i,
μ = fixed overall mean,
t_i = random effect of test year i,
v_j = fixed effect of variety j, and
e_{ij} = random measurement error of variety j in test year i, including any interaction between t_i and v_j.

If all Y_{ij}s are coded by subtracting $(\hat{\mu} + \hat{t}_i)$ and standardized by dividing by s_y, the square root of the mean square error of Y_{ij}, then

$$y_{ij} = \frac{Y_{ij} - \hat{\mu} - \hat{t}_i}{s_y} \approx \frac{v_j + e_{ij}}{s_y}. \quad [1]$$

This revised model differs in three respects from the original 1988 model.

1. As suggested by Piepho, test years are considered random, rather than fixed as before. For the variety testing application test years are considered a form of replication. While varieties under test usually represent a good sample of varieties available to producers in the geographic area for which the testing location corresponds, consideration of varieties as fixed is appropriate because the objective is estimation and comparison of means (Eisenhart, 1947).
2. Standardization is accomplished by dividing by the square root of the mean square error of the overall analysis across test years (s_y) instead of dividing data from each test year by its individual standard error (s_i), derived from the least-significant difference, as before. This solves the problem of improper standardization, which led to errors in probability estimates, as stated by Piepho in the latter portion of the next to last paragraph of his **Problems** section. Note that coding is done as before.
3. Hats are omitted from v_j, e_{ij}, and s_y (formerly s_i) in Eq. [1] because their inclusion before was statistically incorrect.

Because of increased accessibility to automated computation since 1988, i.e., SAS is available on personal computers, and because this revision uses a standardization parameter s_y estimated from a data set combined across test years, calculation of the numerator of the test statistic with least-square means (LSM) is convenient (col. 1, p. 908; Eq. [5], col. 2, p. 910). Unlike the original 1988 method, the combined data set needed to calculate the LSMs has been assembled in order to calculate s_y.

Assumptions

Assumptions are the same as in 1988 except for the additional assumption that the e_{ij}s (errors of all combinations of test years and varieties) are identically normally distributed with mean = 0 and variance = σ^2, estimated by s_y. This differs from before when only the error terms within a test year were required to be identical. As before, we assume cov $(e_{ij}, e_{ij'}) = 0$ but allow cov $(e_{ij}, e_{i'j})$ to vary from zero. Thus the variance-covariance matrix of $(e_{ij}, e_{i'j})$ consists of diagonal elements $(i = i')$ equal to one, because of the assumption of equal variances of test years and standardization with s_y calculated from the combined data set, and off-diagonal elements $(i \neq i')$ which may or may not equal zero.

This model retains the feature of a single error term σ^2, which Piepho expresses as $\sigma^2_{\eta} + \sigma^2_{\varepsilon}$. It is emphasized that the presence of the σ^2_{η} term is recognized, but it is not separated from σ^2_{ε}. Changing the standardization to s_y, the square root of the mean square error of a standard ANOVA table, accomplishes two things. First, it utilizes all available information regarding variation of the error term. Thus a more stable

estimate is obtained. Second, the estimate in this case is an unbiased estimate of

$$\sigma_{iv}^2 + \sigma_c^2.$$

Thus it no longer underestimates error variance. This change effectively eliminates underestimation of the variability of coded values as pointed out by Piepho. Derivation of this result is available upon request from the authors.

Comment 1

Piepho is correct in pointing out that the term σ_v^2, which appears in the numerator of the equation near the center of column 1, p. 911 of the 1988 paper, should be deleted. As he states in the last paragraph of the **Problems** section of his letter, for a given variety j, the effect v_j does not change with i. Thus it has no variance. Given this fact and the new assumption made in the revision,

$$r_{ii'} = cov\,(y_{ij}, y_{i'j}) = \text{cov}\left(\frac{v_j + e_{ij}}{s_y}, \frac{v_j + e_{i'j}}{s_y}\right) = \frac{\text{cov}\,(e_{ij}, e_{i'j})}{s_y^2}, \qquad [2]$$

instead of

$$\frac{\sigma_v^2 + \text{cov}\,(e_{ij}, e_{i'j})}{s_i s_{i'}}$$

as before.

Comment 2

Regarding Piepho's final expression of the variance of the difference between two means (Eq. [3]), namely

$$\text{var}\,(\bar{X}_{.j} - \bar{X}_{.c}) = \frac{\sigma_t^2 + \sigma^2}{n_j} + \frac{\sigma_t^2 + \sigma^2}{n_c} - \frac{2n_{jc}\sigma_t^2}{n_j n_c},$$

if $\sigma_t^2 = 0$ because data are coded, $\sigma^2 = 1$ because data are standardized, ys are substituted for Xs in recognition of coding and standardization, and $n_{j'}$ is substituted for n_c, then this expression becomes

$$\text{var}(\bar{y}_{.j} - \bar{y}_{.j'}) = \frac{1}{n_j} + \frac{1}{n_{j'}},$$

the expression in the 1988 paper (col. 1, p. 912).

This expression was derived from Eq. [10] (col. 2, p. 911) under the additional assumption, cov $(e_{ij}, e_{i'j}) = 0$. Thus Eq. [10] and Piepho's Eq. [3] differ only by the allowance for the covariance term to vary from zero.

As noted by both Piepho in his letter and Jones in the **Reduction to a Familiar Expression** section at the end of the 1988 manuscript, if the two ns are equal and cov $(e_{ij}, e_{i'j}) = 0$, the variance of the difference between two means becomes $2\sigma^2/n$, a familiar expression. Both the revision here and the WLS (weighted least-squares) procedure of Piepho assume homogeneity of error variances across test years. But unlike Piepho's Eq. [3], the probability method requires no estimation of variance components because data have been coded and standardized, simplifying the calculation.

Comment 3

The 1988 paper was directed to the analysis of what is commonly referred to in this country as variety test data, namely comparison of released varieties in tests conducted by state agriculture experiment stations and financed by fees paid for by the proprietary firms whose released varieties are being evaluated. This is made clear in the introduction (p. 907) by a thorough discussion of variety testing by the 12 state agriculture experiment stations which participate in the Central Alfalfa Improvement Conference. In addition, the example presented in Tables 1 and 2 (p. 908) and Fig. 1 (p. 909) was calculated from alfalfa variety test data collected by the Iowa State Agriculture Experiment Station.

In contrast, Piepho's letter regards testing of breeder's lines, a component of a breeding program. This is also the context of the Bradley et al. (1988) paper, which discusses the paired t test, recommended by Piepho as a preferred alternative. A genuine difference in objectives and approach exists between the in-house testing of lines before release, a research activity, and variety tests conducted by state agriculture experiment stations, an extension activity. Again, Piepho's letter and Bradley et al.'s paper pertain to the former situation, while Jones' paper pertains to the latter.

The objective of in-house testing is to identify the breeder or proprietary firm's best material for potential release. As stated by Piepho, time is of the essence so superior materials may be introduced as quickly as possible. Decisions are made based on data collected at a multiplicity of locations to help achieve broad adaptation. The objective of variety tests conducted by state agriculture experiment stations is altogether different. Insofar as possible, a testing location utilized by the agriculture experiment station has been chosen to represent a reasonably well defined intrastate geographical region's climatic and edaphic features. This is tantamount to the statement in the first paragraph of the introduction of the 1988 paper, "Combining results across tests conducted at a single location or across locations known not to exhibit significant amounts of nonrandom, predictable genotype × location interaction is desirable."

This is a logical approach for varieties already available for sale to the public because the user of the extension variety test report is interested in how well a certain variety is likely to perform at his location. Information detailed in these reports is generally considered more reliable than data offered by the proprietary firms themselves because it is assembled by the presumably unbiased state agriculture experiment station. Because this approach has worked so well in the past, it is unlikely to be substantially altered in the foreseeable future.

One to several locations may be utilized in a state, but test results are reported separately for each location(s) corresponding to the above-described geographical region. Of greatest interest is combining test-year data within the geographical region. The intent of the 1988 paper was to describe a method to accomplish this goal. The criticism that the method is irrelevant because the number of years is prohibitive is meaningless because the material being tested by the agriculture experiment station is already available for sale to the public. Furthermore, the method showed in the alfalfa example that only seven test years were necessary to evaluate these varieties rather than the 12 to 23 test years of data accumulated for the 15 proprietary varieties. Another point to be considered is that these multiple test years need not be sequential. They may overlap, as described for the alfalfa example in the first paragraph of **Materials and Methods** (Jones, 1988). The same approach could be applied with annual crops by varying date of planting or some other management variable appropriate to the geographical region. For either annuals or perennials, additional locations representative of the geographical region could be used.

A further issue concerns Shukla's (1972) stability parameter. We assume that the e_{ij}s are identically normally distributed with mean = 0 and variance = $\sigma_{iv}^2 + \sigma_\varepsilon^2$, regardless of variety. As Piepho has pointed out, this is incompatible with the Shukla

stability parameter concept. However, the goal of variety testing and the probability method is to distinguish between performance of released varieties for a reasonably well defined geographical area. From this point of view, stability parameters do not hold the intrinsic importance they would if the objective were to evaluate breeder's lines over a wide geographical area. The test year functions as a form of replication in variety testing as well as an environment in the sense considered in the stability parameter literature. Therefore, in our minds it is logical to retain this assumption for the variety testing application.

Discussion of the Three Methods

1. Piepho's WLS Procedure

This procedure assumes the e_{ij}s are identically normally distributed with mean = 0 and variance = σ^2. If the weights have large variances, WLS can cause very large errors in estimates. It is commonly accepted that a somewhat biased estimate with a small variance is preferred over a less biased estimate with large variance. In practical variety testing situations, acceptably accurate weights will generally be unavailable because of insufficient sample size. Assuming cov (e_{ij}, $e_{i'j}$) = 0 as Piepho suggests still requires estimation of σ_i^2 and σ^2, despite the fact that Eq. [3] is structurally analogous to the 1988 Eq. [10], which requires no variance component estimation, as detailed in **Comment 2**.

2. Paired *t* Test

This approach is obvious. It has the advantage of being easily applied statistically because of its simplicity. Unfortunately, this same feature requires much of the data in unbalanced data sets to be excluded. As a procedure, then, it puts us firmly back at the proverbial square one as far as dealing with the challenge of unbalanced data in variety testing.

3. Revised Probability Method

The revision, as opposed to the method as first proposed (Jones, 1988), requires the assumption of a homogeneous σ^2 across test years and varieties. This permits standardization with the common σ^2, but at the same time downplays the importance of variation in stability of varieties, a disadvantage in the eyes of some. However, the revised probability method permits cov (e_{ij}, $e_{i'j}$) to differ from zero without requiring statistical exercises in estimation, as does Piepho's WLS procedure. This revision has the advantage over the paired *t* test of permitting inclusion of all available data in calculation of the test statistic.

USDA-ARS, THOMAS A. JONES
Forage and Range Res. Lab.
Utah State Univ., Logan, UT 84322-6300.

Dept. of Mathematics and Statistics, RONALD V. CANFIELD
Utah State Univ., Logan, UT 84322-3900.

References

Bradley, J.P., K.H. Knittle, and A.F. Troyer. 1988. Statistical methods in seed corn product selection. J. Prod. Agric. 1:34–38.

Eisenhart, C. 1947. The assumptions underlying the analysis of variance. Biometrics 3:1–21.

Jones, T.A. 1988. A probability method for comparing varieties against checks. Crop Sci. 28:907–912.

Shukla, G.K. 1972. Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29:237–245.